Evidence & Policy • vol xx • no xx • 1–27 • © Policy Press 2021
Print ISSN 1744-2648 • Online ISSN 1744-2656 • https://doi.org/10.1332/174426421X16141794148067
Accepted for publication 24 February 2021 • First published online 29 March 2021

research

Does evaluation quality enhance evaluation use?

Pirmin Bundi, pirmin.bundi@unil.ch
University of Lausanne, Switzerland

Kathrin Frey, frey@kek.ch
KEK-CDC Consultants, Zurich, Switzerland

Thomas Widmer, thow@ipz.uzh.ch
University of Zurich, Switzerland

Background: Evaluations are a useful tool to learn more about the effectiveness of public measures. In the era of evidence-based policymaking, recent research suggests that quality is an important determinant of the utilisation of evaluations. Despite this claim, hardly any empirical study has investigated whether the quality of an evaluation – measured by a meta-evaluation – influences its perceived utilisation by decision makers.
Aims and objectives: This article asks how the quality of an evaluation study is related to its perceived use, and investigates the relationship between the quality of an evaluation, assessed through a meta-evaluation, and how the evaluation is perceived and accepted by the parties concerned.
Methods: The basis for the empirical analyses were 34 external evaluations, conducted from 2006 to 2014, of upper secondary schools in the canton of Zurich, as well as a standardised survey conducted among 307 representatives of these schools (teachers, administrators, members of quality development teams, and the heads of school oversight commissions).
Findings: We conclude that the quality of the evaluation, as assessed in a meta-evaluation, is not particularly associated with the perception of evaluation quality and the perceived use of the evaluation. The perceived quality, however, is related to the perceived impact of an evaluation.
Discussion and conclusion: These findings are relevant for evaluation research and practice, since they show that the quality of an evaluation and evaluation use do not necessarily go hand in hand.

Key words evaluation quality • evaluation use • meta-evaluation • survey

Key messages

- Evaluators have to be aware that a systematically assessed quality of an evaluation does not go hand in hand with the perceived quality of that evaluation;
- Evaluators often focus on the instrumental form of evaluation use, but they should not ignore other forms of use and maybe try to maximise these utilisation forms in the design of their evaluation;
- Evaluators should be more active in advising stakeholders when it comes to evaluation use, for example, through policy narratives;
- Evaluators should carefully think about the measurement of evaluation quality and evaluation effects in research on evaluation.

To cite this article: Bundi, P., Frey, K. and Widmer, T. (2021) Does evaluation quality enhance evaluation use?, *Evidence & Policy*, vol xx, no xx, 1–27,
DOI: 10.1332/174426421X16141794148067

Introduction

Public authorities continue to use evaluation systems to govern partly autonomous public or private institutions, as evaluations provide decision makers with information on the effectiveness, efficiency or relevance of public actions using scientific methods (Gaertner et al, 2014; Verhoest et al, 2007; Ehren et al, 2015).[1] They routinely commission evaluations in order to assess how public or private institutions carry out the public services assigned to them. The intent is to oversee and improve what such institutions provide, and to do so using evaluation findings (Bundi, 2016). This is particularly true in the education sector, where school evaluations have long been a central element for ensuring the success of a curriculum (House, 1993; Cronbach, 2000).[2] Evaluation findings have also become a source of information for decisions about the educational system as a whole and have become relevant for individual schools as well (Sanders and Horn, 1998; Odom et al, 2005; Slavin, 2008; Seidel et al, 2017).

The assumption that evaluations provide high-quality evidence legitimises them in the eyes of the public authorities. Crucial elements here are the methodological quality of the evaluations themselves, together with, inter alia, the stakeholder support for the evaluation process as well as the trustworthiness of the evaluation findings, not just by the public authorities but also by the evaluated schools. Public authorities that introduce evaluation schemes are generally interested in high-quality evaluations in order to legitimate their use and to allow installing subsequently a mechanism to increase their use. In this context, the question whether evaluation quality – and which dimensions of it – affects the use of evaluations becomes essential.

There is a rich research literature on the use of evaluations and their influence that has demonstrated that evaluation quality indeed affects evaluation use (Cousins and Leithwood, 1986; Johnson et al, 2009). One of the most common distinctions conceives four different types of use (Alkin and King, 2016). Instrumental use refers to the direct use of systematically generated knowledge (for example, evaluations) to take action or make decisions. Conceptual use points at indirect use of systematically generated knowledge that opens up new ways of thinking and understanding, or that generates new attitudes or changes existing ones. Thirdly, one can distinguish symbolic use which refers to the use of evaluations to support an already preconceived position in order to legitimise, justify or convince others of their position. Finally, process use refers explicitly to use that occurs due to the process and not due to the results of an evaluation. According to Patton (1997: 90), process use has developed as a term to address the way in which activities that occurred during an evaluation affected individuals or an organisation, in contrast to the results of an evaluation. In doing so, process use is characterised as the individual change in thinking and behaviour that occurs to those involved in the evaluation through learning during the evaluation process.

However, most studies on evaluation quality have not distinguished between different types of evaluation and more importantly, they rely on the quality of the evaluation as perceived by the involved actors rather than on a systematic external quality assessment using evaluation standards (Dedering and Mueller, 2011; Ledermann, 2012; Gaertner et al, 2014; Böhm-Kasper et al, 2016). Albeit previous studies show that stakeholder involvement (Ayers, 1987; Cousins, 1995; Earl, 1995; Lafleur, 1995; Lee and Cousins, 1995; Turnbull, 1999; Johnson et al, 2009); the communication of the evaluation results (Cousins and Leithwood, 1986; Marsh and Glassick, 1988; Lafleur, 1995); the use of mixed methods (Potts, 1998); and the political context (Newman et al, 1987; Weiss et al, 2005) all foster the use of evaluation, there is little information available on how, or if, the results of an external quality assessment of evaluations relates to the perceived evaluation quality and acceptance, as well as about the consequences this may have for the use and influence of evaluations. An external assessment utilises a set of predefined standards of evaluation quality. These standards always involve trade-offs, so it becomes relevant to investigate the influence of the individual criteria used in these standards.

This study examines the relationship between evaluation quality as measured through a meta-evaluation and the implementation of the evaluation findings and their impact.[3] Based on the influence model of Henry and Mark (2003), we argue that the quality measured by an external meta-evaluation can provide information in order to understand differences in the perceived quality and acceptance of an evaluation at the school officials level, and in the (perceived) evaluation consequences at the collective level of the partly autonomous public or private schools.

The empirical analysis uses 34 external evaluations of upper secondary schools in the canton of Zurich. The Zurich Department of Education, which has conducted external evaluations of upper secondary schools since 2005, commissioned the Institute for External School Evaluation (IFES)[4] to conduct external evaluations of the 34 upper secondary schools in the canton of Zurich. The aims of these evaluations were to assist individual schools in improving their quality by providing stimuli for school development, by providing relevant information for administering and overseeing the upper secondary schools in the canton, and to improve the political and public accountability of the schools. The 34 schools were subjected to a meta-evaluation of their quality based on a subset of the criteria listed in the 2001 Swiss Evaluation Society (SEVAL) standards (Widmer et al, 2000), as well as customised criteria relevant to the specific setting (such as coverage of leading questions or needs satisfaction). Moreover, a survey among those who 'represent' the school collected data on their views of evaluation processes and results, as well as about the perceived use and impact of the evaluation findings within the involved schools.

Our findings suggest that the external assessment of quality needs to be differentiated and has a complex relationship to the perceived quality and acceptance of the evaluation. While some dimensions appear to be related to the perceived evaluation quality and acceptance, other dimensions do not correlate with it. Moreover, the perceived implementation of evaluation results is not necessarily linked to a higher perceived evaluation impact. In contrast, the perceived evaluation quality correlates significantly with the perceived evaluation impact. Hence, the findings of the study suggest that the evaluation findings were not used instrumentally, but other types of use such as process or conceptual use seem plausible, which might be avenues of future research.

The article is structured as follows. The first section introduces the concept of evaluation quality and highlights its different dimensions. It discusses how evaluation quality is related to evaluation use and impact and presents the analytical framework that we will empirically test. The following section introduces the evaluation system used in the governance of upper secondary schools in the canton of Zurich. After a section about methods and data, the results section follows. The last sections discuss the results and point out the relevance and implications for evaluation systems and research on evaluation.

Literature review

Evaluation quality

Since evaluations are often undertaken under the constraints of time, money, and political pressures, there have long been concerns expressed in the literature about the resulting quality (for early examples, see Stufflebeam, 1974; Cook and Gruder, 1978 (esp. 15–16); Hatry, 1980; House, 1980). One part of this discussion deals with assessing the worth and merit of an evaluation using meta-evaluative studies (Scriven, 1969; Stufflebeam, 1974; 2001; Cook and Gruder, 1978; Widmer, 1996; Schwartz and Mayne, 2005; Cooksy and Caracelli, 2009). Meta-evaluations rely on the assumption that 'evaluation systems can be considered as social programs' (Cook and Shadish, 1982: 232) and therefore are evaluation objects as social programmes. This kind of 'second order evaluation' (Scriven, 1969: 36) requires assessment criteria for ensuring a transparent and systematic assessment of the meta-evaluated object.

A more general discussion about evaluation quality beyond meta-evaluations started in the 1970s and produced several sets of criteria for assessing the quality of an evaluation. These sets, often called 'evaluation standards', include the desired qualities evaluation studies should possess. The most significant among them would be the Standards for Evaluation of Educational Programs, Projects, and Materials, edited by the Joint Committee on Standards for Educational Evaluation (1981). These 'Program Evaluation Standards', as they were called after the publication of a second edition (Joint Committee on Standards for Educational Evaluation, 1994), were disseminated widely and were used by several national evaluation societies as a basis for developing their own set of standards (Russon and Russon, 2004; Widmer, 2004). The Swiss Evaluation Society (SEVAL) was the first among them, adopting the SEVAL Standards in 2001 (Widmer et al, 2000). The SEVAL Standards postulate – as do many other national standards of this kind – that an evaluation needs to account for four features: utility, feasibility, propriety, and accuracy. The standards thus define the quality of an evaluation by taking into account, as far as possible, the four characteristics simultaneously. Each of these dimensions contains individual standards designed to reflect the particular feature. The SEVAL Standards includes 27 such items that serve as an instrument for measuring evaluation quality.

In addition to these general requirements for establishing the quality of an evaluation, which are also called external assessment criteria and that are regarded as valid for any type of evaluation, there are also internal assessment criteria. These are specific expectations, such as the objectives of a given evaluation, and they are relevant for a comprehensive and balanced assessment of that evaluation (Widmer, 2005: 49). These specifications may be laid out in the documents (such as in the requests for proposals

or evaluation plans), but they may also come from stakeholder groups articulating the particular expectations they have of an evaluation (such as the properties of deliverables, definitions of the evaluation criteria to be used, target groups who are to be involved, or the coverage of select evaluation questions). The general (external) and the specific (internal) expectations may well overlap, or the latter may be an operationalisation of the former.

The literature emphasises that the procedure for selecting evaluation criteria is a crucial point a meta-evaluation should focus on. One comparative analysis argues that 'the use of some tailored or pre-specified quality criteria could provide a stronger basis for conclusions about the evaluation's quality' (Cooksy and Caracelli, 2009: 10), whereas 'a goal-free meta-evaluation, using emergent criteria…, can provide a full range of strengths and weaknesses that may be useful in improving a final report' (Cooksy and Caracelli, 2009: 10) in the context of a formative meta-evaluation (Bustelo, 2002: 4, 6). The question of evaluation quality has also received particular attention in the debate on evidence-based policymaking (Davies et al, 2000; Nutley et al, 2003; Pawson, 2006); at issue is here which findings count as good (or rigorous) enough evidence (Donaldson et al, 2016).

Evaluation use and influence

Mark and Henry (2004) have proposed a sophisticated theory of evaluation influence which differentiates between evaluation inputs, evaluation activities, evaluation outputs, intermediate and long-term evaluation outcomes, and contextual factors. They also distinguish between individual, interpersonal, and collective levels, and specify mechanisms at each level for how evaluation can support social improvement (change). These mechanisms capture underlying processes by which evaluations influence attitudes, motivations and actions.

The present study applies this model of change to the context of evaluating partly autonomous public and private schools. It takes up the core expectations concerning the consequences of evaluation and the assumptions, both in the literature and in practice, of how such systems of evaluation should lead to these consequences (Petrosino et al, 2005; Ehren and Visscher, 2006; Verhoest et al, 2007; King and Alkin, 2019; Head, 2010). It defines the evaluation activities, outputs, and the intermediate and long-term outcomes of interest. Moreover, the use of evaluations is strongly linked to the expectation that they will lead to improvements in the evaluated schools. The use of an evaluation is therefore expected to lead to immediate, specific, purposeful activities (for example, measures), or in other words, to an instrumental use of evaluation (see Cousins and Leithwood, 1986; Weiss, 1998).

In the light of Mark and Henry's (2004) theory, we therefore focused on the intended behavioural processes at the collective school level as intermediate evaluation outcomes. To support such instrumental use, public authorities can introduce incentives (rewards and sanctions) or impose obligations that require the implementation of evaluation-based measures (see Verhoest et al, 2007; Ehren et al, 2015). These measures aim for long-term improvements in these schools. For public schools, the expectation is that evaluation-based measures foster improvements in key areas of school governance and success in areas such as student performance, school atmosphere, teaching quality, or quality management (Ehren and Visscher, 2006; Böttcher and Keune, 2010; Quesel et al, 2011a). Any of these evaluation outcomes

can originate in the information-processing behaviour of the actors involved (Mark and Henry, 2004). Evaluation systems assume that the people in charge of school governance process the information provided by the evaluation, and thereupon draw up measures to be implemented which will improve their schools.

While Mark and Henry (2004) lay out the full range of cognitive, affective, and motivational processes that can lead to the intended behaviours, our focus here was on a key assumption of external evaluation systems: that evaluations are more likely to trigger these processes and changes if they provide evaluative information of high quality. Hence, we focused on one particular pathway Mark and Henry (2004) identify, namely the relationship between the quality of an evaluation, as externally assessed, the perception of the quality and acceptance of the evaluation by the individual school representatives involved, the perception of the implementation of evaluation-based measures, and the impact the evaluation has on key areas of performance in these schools.

Put more simply, our study asks whether evaluations which are assessed – in a systematic meta-evaluation – as being of higher quality are also perceived – in a field (or school) setting – as being of higher quality and thus more accepted. Do they then also generate more advanced perceived implementation of evaluation-based measures within the evaluated school? Hence, we argue that the advanced perceived implementation of evaluation-based measures should lead to measurable improvements of the schools.

As Rickinson et al (2020) rightly emphasise, it is not only important whether evidence from evaluations is used, but also how the evidence is used. Our study reaches its limits here and cannot consider this very important aspect of the quality of evidence use.

Analytical framework

In research on evaluation, the perception of the quality of an evaluation has widely been used as an explanatory factor for the use and influence of evaluations (Weiss and Bucuvalas, 1980; Cousins and Leithwood, 1986; Johnson et al, 2009). Perceptions of quality can refer to the users' judgement of an evaluator, the evaluation process, and results (for example, as set out in a report) in terms of accuracy or credibility (Mark and Henry, 2004; Miller, 2015). Most empirical studies show that the perception of quality does matter. However, research remains inconclusive as to its importance relative to other factors, which are associated with the use and influence of an evaluation (Johnson et al, 2009; Frey, 2010; Ledermann, 2012).

By contrast, research on evaluation has rarely tried to explain the perception of quality (Miller, 2015). The rare empirical studies tend to focus on a particular dimension of quality, such as evaluation methods (Jacobson and Azzam, 2016), or on the relationship between the content of the evaluation findings and the perceived quality of the evaluation (Jacobson and Azzam, 2016; Weiss et al, 2008). Miller (2015), with reference to the broader literature on individual information processing and credibility judgements, emphasised that factors beyond evaluation methods and content determine whether the findings of an evaluation are considered credible.

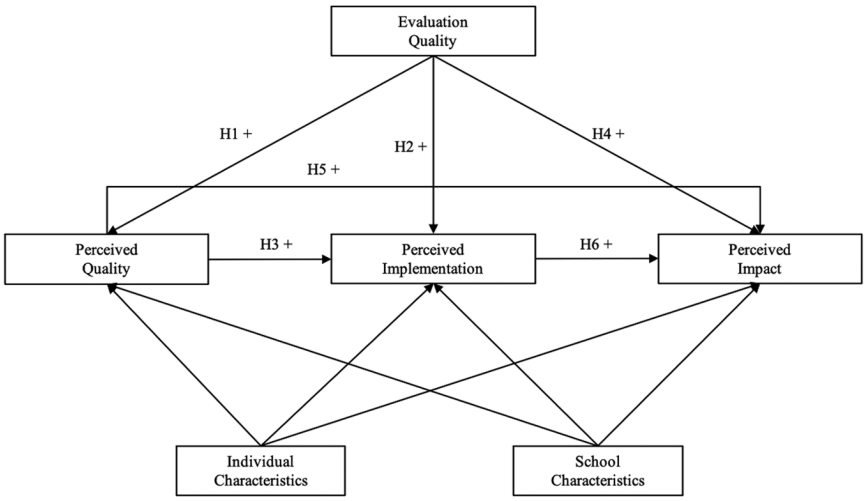
The literature convincingly argues that the influence of evaluation aspects on the perception of quality differs by the characteristics of the users of an evaluation and the context in which it is conducted or applied (Mark and Henry, 2004). The relevant

framework in the context of our analysis, for example, includes characteristics of the schools evaluated and the individual users. In doing so, they can be considered as additional exogenous variables. In addition, how involved those who will use or be affected by an evaluation is an important additional aspect, and reviews of the empirical literature (Johnson et al, 2009: 389; Daigneault, 2014) indicate that stakeholder involvement is a key element in fostering the use of evaluation. Therefore, the involvement of representatives in the evaluation is linked with *their perception* of the evaluation quality and impact.

Our theoretical framework suggests that the quality as systematically assessed in a meta-evaluation is related to the perceived quality of an evaluation as well as its perceived acceptance. The higher the quality is assessed in a meta-evaluation, the more likely a member of the evaluated school will not only perceive the evaluation of high quality, but also accept it. As outlined above, a meta-evaluation involves multi-dimensional quality criteria. A blanket assessment of evaluation quality is not appropriate because of the multidimensional nature of quality. In addition, a cross-dimensional assessment would ignore that not every criterion should have the same importance or weight. Therefore, the analytical framework focuses on individual criteria. We argue that an evaluation which takes a predefined quality criterion – such as a general evaluation standard or a customised quality requirement – into consideration is generally more persuasive and perceived as more credible than an evaluation which does not do so. Further, the framework postulates that the quality assessed in meta-evaluation correlates with the acceptance of an evaluation. Oliver et al (2020: 12) show that a stakeholder will more likely accept an evaluation if they share criteria of credible evidence. Moreover, Penninckx et al (2016: 338) argue that perceived consideration of quality criteria, such as transparency, can increase schools' satisfaction with external evaluations as well as their willingness to accept them. Substantially, if stakeholders accept the evaluation results, we expect a higher perceived implementation of evaluation-based measures and the perceived impact. According to Ouimet et al (2009: 338), the adoption and implementation of measures refers to the impact of an evaluation. And, the more advanced the perceived implementation of the evaluation-based measure, the more likely stakeholders will perceive an impact of the evaluation on key performance dimensions. Hence, we postulate the following hypotheses that we are going to test empirically (see Figure 1):

H1: The higher the quality is assessed in a meta-evaluation, the higher a representative of the evaluated school will perceive the evaluation of high quality.
H2: The higher the quality is assessed in a meta-evaluation, the stronger a representative of the evaluated school will perceive the implementation of evaluation findings.
H3: The higher the quality is perceived by a representative, the stronger a representative of the evaluated school will perceive the implementation of evaluation findings.
H4: The higher the quality is assessed in a meta-evaluation, the higher a representative of the evaluated school will perceive the impact of an evaluation.
H5: The higher the quality is perceived by a representative, the higher a representative of the evaluated school will perceive the impact of an evaluation.
H6: The higher the implementation is perceived by a representative, the higher a representative of the evaluated school will perceive the impact of an evaluation.

Figure 1: Theoretical model



Methods and data

The empirical analysis is based on external evaluations of partly autonomous public and private upper secondary schools in the canton of Zurich (Widmer et al, 2015). Educational administration in Switzerland nowadays involves widespread use of external evaluations, but this development has involved a trade-off. As in many countries, reforms of educational institutions have involved increasing autonomy, moving away from the tutelage of the state, and giving schools more freedom to formulate policy and conduct their own internal affairs (Ehren et al, 2015; Gaertner et al, 2014). This autonomy was granted under the premise that it would lead to better performance, with education provided more effectively and efficiently. In return, regular external evaluations were introduced as an instrument to hold the school accountable and to guarantee minimum norms were being upheld. However, external evaluations were not introduced only for accountability purposes, but also as an instrument for further developing and improving the schools (Gaertner et al, 2014).

In order to study the relationship between the quality of an evaluation and the perceived use of an evaluation in the upper secondary schools, we gathered data from two different sources – a meta-evaluation and a survey among school representatives that were both conducted during an external evaluation of the evaluation system. A list of 36 quality assessment criteria was drafted based on the SEVAL Standards (Widmer et al, 2000), reference documents from the cantonal Department of Education, and from the IFES.[5] The original 36 criteria were grouped into five categories: utility, feasibility, propriety, accuracy, and customised criteria.[6] However, 15 of these criteria could not be included, as they could not be assessed based only on an analysis of the written reports, which is why we use only a subset of the SEVAL criteria. As a result, the study conducted a qualitative content analysis (Mayring, 2015) using 21 criteria; nine of them showed no variance among the reports and were therefore dropped from the analysis. This result indicates that the evaluations assessed in this meta-evaluation share certain similarities, facilitating a comparative meta-evaluation (Bustelo, 2002: 14).

To assess the reliability of the coding, two persons independently coded the evaluation reports on a five-point scale from 'negative' to 'positive'. Indicators were established that had to be found in the evaluation reports in order to assess whether or not the criteria were met.[7] The coding achieved a considerable intercoder reliability,[8] and was discussed and reconciled within the research team. Moreover, in order to compare the scores of the ratings with each other, the coefficients have been standardised. In doing so, the values have been rescaled to a standard deviation of 1 and a mean of 0.

The meta-evaluation of the external evaluations used the most recent written evaluation reports about each school at the time of our study, so of the 34 schools evaluated, 21 were being externally evaluated for the very first time and the rest were in the second cycle of external school evaluation. The design of these evaluations was only marginally modified after the first cycle, so there was little systematic difference in the first and second evaluation cycle reports.

This study also employed data from an online survey among school representatives from all 34 secondary schools during October and November 2014. The survey focused on experiences with the most recent external evaluation of each school, so we were able to match the data from the meta-evaluation with the data from this survey. We surveyed members of the respective school administrations, the school quality development teams, teachers, and the presidents of the oversight commission of each school. As in some cases external evaluations were conducted several years previously, there was an issue with personnel turnover. If possible, we invited responses not only from the current but also from the former holders of relevant positions during the external school evaluation. In total, 307 representatives of the schools, including 121 members of the school administrations, 112 members of the school quality development teams, 37 teachers, and 37 presidents of the oversight commissions participated in the survey. This corresponds to a high response rate of 74.7%.

The empirical analysis includes five endogenous variables. The first two variables, the *perceived evaluation process* and the *perceived evaluation acceptance* are composed by twelve (process) resp. eight (acceptance) different indicators that we have combined in an index.[9] In addition, we have combined both variables into a new variable called *perceived evaluation quality*.[10] The fourth variable asked representatives of the school whether they *implemented measures* they had negotiated with the school authority. The measures are discussed with the cantonal Department of Education and the respective school management and are based on the recommendations of the evaluation report. Moreover, they are highly present in the school directions' consideration in the aftermath of the evaluation. In doing so, this type of use can be characterised as instrumental use, which refers to the direct use of systematically generated knowledge (for example, evaluation findings) to take action or make decisions (Alkin and King, 2016).

The fifth variable, the *perceived evaluation impact*, was operationalised using 13 different impact dimensions (for example, students' performance, school atmosphere, teaching atmosphere, teaching quality) developed by Quesel et al (2011b) to measure the impact of external school evaluations on primary schools.[11] Some of the 13 dimensions of impact included (such as the impact on the school quality management) could qualify as well as intermediate outcomes, and not as a final impact of interest (such as students' performance). Effects of this kind, depending on the perspective, can also be seen as a value in themselves, sufficiently justifying their inclusion. The school's representatives

were asked whether the external school evaluation had a strong positive, a positive, none, a negative, or a strong negative impact on these dimensions.[12]

Several other individual, and institutional, exogenous variables were also integrated into the empirical analysis in order to control for the main correlations. Involvement in the evaluation was measured with three dimensions. School representatives were asked to what extent they could influence the evaluation instruments, the design of the on-site visits, and the formulation of recommendations. The year of evaluation was included to measure the time that had passed since the evaluation was carried out. Another substantial constraint of the study is the retrospective character of the survey respondents' ratings of the evaluation's impact and the perceived implementation of the evaluation recommendations. It is plausible that perceptions of quality might be affected by the perceived evaluation impact, and by including the time span we can partly account for this limitation. The variables, gender of the interviewed person and school type (academic high school or vocational school) were measured with dummy variables.

Our observations are nested in groups (schools) that have the potential to influence the quality and perceived use of evaluations. According to Steenbergen and Jones (2002: 219–220), ignoring the clustering of the data structure could lead to biased standard errors that would overestimate the significance of effects. However, since we are not interested in any factors on the school level, we use clustered standard errors (Rogers, 1993). Robust variance estimation allows not only for the relaxation of the assumption that the error terms are identically distributed, but also clustering allows the further relaxation of the assumption that the observations are fully independent.

Findings

Table 1 presents the findings of the regression models to explain the perceived quality, implementation and impact of the evaluation.

First of all, Model 1 suggest that the quality of the evaluation, as assessed in a meta-evaluation, is not positively related to the perceived evaluation quality. In doing so, only the used method seems to be slightly associated with the perceived evaluation quality: If the evaluation contains qualitative and quantitative data collection and analyses, then the school representatives accept the evaluation results less often. As a consequence, we have to reject our first hypothesis. In contrast, women and representatives who were more involved in the evaluation have a more positive view of its quality. The latter is often argued in the studies on evaluation use (Cousins and Leithwood, 1986; Johnson et al, 2009; Daigneault, 2014, and so on). Furthermore, representatives from vocational schools are more likely to accept the evaluation, and the less recently an evaluation was conducted, the more negative the perceived acceptance of the evaluation.[13]

As a next step, we analyse whether the measured evaluation quality influences the perceived implementation of the evaluation (Model 2). On the one hand, the perceived implementation is only related to the precise description of the evaluation object and the needs orientation. The better the description and the analysis of the context, the more frequently measures were implemented which the cantonal authorities stipulated. Moreover, the more the evaluation report focused on the context and the needs of the school (IFES category), the stronger the evaluation implementation. On the contrary,

the less neutral the evaluation report and the fewer the utilisation of qualitative and quantitative analyses, the more frequently the implementation of measures. The same is true for the coverage of central questions. Since the Department of Education demands to implement the stipulated measures, we are not surprised about that result.

Table 1: Meta-evaluation, perceived evaluation quality, implantation and impact[14]

	Perceived quality	Perceived implementation	Perceived impact
	(1)	(2)	(3)
Evaluation quality in meta-evaluation			
Subset of SEVAL standards			
Identifying stakeholders	–0.003 (0.179)	0.078 (0.166)	0.121*** (0.038)
Transparency of value judgements	0.079 (0.113)	–0.129 (0.155)	–0.037 (0.036)
Comprehensiveness in reporting	0.116 (0.147)	-0.077 (0.197)	-0.020 (0.038)
Complete and balanced assessment	–0.013 (0.079)	–0.002 (0.090)	0.018 (0.022)
Precise description of the evaluation object	0.071 (0.080)	0.238** (0.110)	0.049* (0.025)
Analysing the context	0.017 (0.072)	0.200* (0.101)	0.039* (0.020)
Goals, questions and procedures	0.082 (0.056)	–0.171 (0.101)	0.030 (0.055)
Qualitative and quantitative analysis	–0.131* (0.071)	–0.212* (0.110)	0.086*** (0.024)
Substantiated conclusions	–0.099 (0.077)	0.070 (0.108)	0.045* (0.072)
Neutral reporting	–0.016 (0.140)	–0.368*** (0.129)	–0.081*** (0.028)
Subset of canton Zurich criteria			
Coverage of central questions	–0.117 (0.077)	–0.321** (0.122)	–0.046* (0.026)
Subset of IFES criteria			
Need orientation	–0.047 (0.180)	0.464* (0.268)	–0.075* (0.042)
Individual characteristics			
Woman	0.242*** (0.077)	0.196 (0.229)	–0.008 (0.062)
Involvement in the evaluation	0.240** (0.096)	0.150 (0.232)	0.090 (0.063)
School characteristics			
Vocational school	0.277* (0.146)	0.271 (0.212)	0.103** (0.048)
Year of evaluation	–0.038 (0.028)	0.021 (0.039)	0.024*** (0.009)
Perceived evaluation quality		0.020 (0.231)	0.497*** (0.051)
Perceived implementation			–0.004 (0.031)
Constant	1.818*** (0.596)	2.931*** (0.893)	1.666*** (0.208)
N	124	92	92
F(Prob > F)	22.84***	2.69***	35.03***
R2	0.388	0.163	0.655

Note: results are from a regression model with school clustered standard errors. Standardised regression coefficients shown with robust standard errors in parentheses. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

It might be that the schools prefer less balanced reports, as they provide them with the possibility to influence the measures more concretely if they are overwhelmingly positive or negative. Moreover, Olsen et al (2013) show that subjects of evaluations are purposively selected following a process that is non-random, which could lead to a bias in the evaluation report. We find therefore evidence for and against our hypothesis 2, which we cannot confirm. However, our model suggests that there is no significant correlation between the perceived evaluation quality and the perceived implementation, which is why we have to reject our hypothesis 3.

Model 3 in Table 1 presents the results for the perceived impact. On the other hand, the meta-evaluation assessment of quality correlates selectively with the evaluation's impact. The better the evaluation report has identified the stakeholders, the higher the respondents have perceived the impact of the evaluation. While the utilisation of mixed methods had a negative effect for the perceived evaluation quality and the implementation, it is associated positively with the impact. In addition, the precise description of the evaluation object, the analysis of the context, and substantial conclusions, are positively related. In contrast to the evaluation implementation, the needs orientation of the evaluation report is negatively related with the impact. Again, our model suggest that the meta-evaluation criteria are not one-dimensionally linked to the perceived impact, which is why we cannot confirm hypothesis 4. Furthermore, the less time has passed since the evaluation was conducted, the higher the perceived impact. Vocational schools tend to perceive the evaluation impact as higher, which correlated with the higher perceived evaluation acceptance by them. Last, and most important, the perceived evaluation quality correlates with the perceived impact, while the perceived implementation of measures is not associated with it. Therefore, we can confirm hypothesis 5, but we have to reject hypothesis 6.

Overall, the analyses illustrate that the meta-evaluation results are mostly not related with the perceived evaluation quality and acceptance, as well as the perceived implementation and impact. Individual and institutional factors seem to matter more than the quality as assessed in a systematic meta-evaluation: women, highly involved school representatives and representatives from vocational schools perceive the quality and the acceptance of an evaluation as higher than other respondents do. This finding corresponds to studies which emphasise the importance of the involvement of individuals and groups invested in the evaluation object (Johnson et al, 2009; Daigneault, 2014). Moreover, the model suggests that the perceived evaluation quality is in fact more important for the acceptance of the evaluation, their implementation and their impact than the quality measured by the meta-evaluation, which partly confirms prior studies (Weiss and Bucuvalas, 1980; and Leithwood, 1986; Johnson et al, 2009). However, we have to emphasise that our regression coefficients are overall rather small and most standard errors large, which is probably due to the small number of observations of our analyses. In contrast, the R-squares are quite large, which point out that our variables are able to explain a lot of variation of our independent variables. Still, our findings should cautiously be interpreted. Table 2 presents a summary of the major results of the study.

Table 2: Summary of the major results of the study

Factors	Perceived quality	Perceived implementation	Perceived impact
Identifying stakeholders	0	0	+
Precise description of the evaluation object	0	+	+
Analysing the context	0	+	+
Qualitative and quantitative analysis	–	–	+
Substantiated conclusions	0	0	+
Neutral reporting	0	–	–
Coverage of central questions	0	–	–
Need orientation	0	+	–
Woman	+	0	0
Involvement in the evaluation	+	0	0
Vocational school	+	0	+
Year of evaluation	0	0	+
Perceived evaluation quality		0	+
Perceived implementation			0

Note: 0: no significant effect; -: negative signficiant effect; +: positive significiant effect

Discussion

Our theoretical framework postulated a positive relationship between the assessed criteria regarding quality and the variables associated with the influence of the evaluation. However, the empirical analysis revealed several negative associations, and most of the findings suggested that the meta-evaluation findings are not related to the perception of the evaluation at all. The questions arise whether (a) an evaluator should pay more attention to certain quality standards over others in the evaluation design to increase the utilisation of an evaluation, and (b) the evaluation community should reconsider the utilisation focus on all evaluation standards.

On the one hand, the meta-evaluation indicated that there are relatively limited differences among the 12 standards considered in the analysis. The school evaluations followed fairly standardised procedures. On the other hand, the analysis draws on only 12 of the 36 standards we developed (based on a subset of the SEVAL Standards and reference documents from the key stakeholders). The analysis excluded 9 Standards due to an absence of variance and 15 Standards due to missing values. In the end, the case selection of a meta-evaluation is always a trade-off between the comparability and the necessity to obtain variance. The IFES evaluation design provided us with a unique opportunity to assess and compare the evaluation quality, but with the cost that the evaluations are quite standardised. We are aware that the scope of the meta-evaluation is limited but, on the other hand, it provides a reliable assessment of comparable evaluations. Hence, we can only make statements about the relevance of certain evaluation standards. Moreover, we might observe a different relationship between a measured and perceived evaluation quality if we used a more limited definition of quality.

Moreover, it is striking that perceived evaluation quality does correlate with the perceived implementation of evaluation-based measures and the perceived impact of

an evaluation, but the implementation of measures is not associated with the perceived impact. This finding suggests that the perceived evaluation quality can induce a non-measurable relation: school representatives are more likely to perceive an evaluation as effective in improving key areas of the school if they perceive the quality of the evaluation as high. In other words, the perceived implementation of evaluation-based measures does not play a role in the perception of impact. As public authorities negotiate with every school about implementing evaluation-based measures, school representatives might perceive these measures as externally determined tasks they need to implement, regardless of the quality of an evaluation.

In order to understand the relationship between evaluation quality and evaluation use, we have assumed that there has to be an instrumental form of use, since the schools stipulate measures with the cantonal department. However, our empirical analysis cannot provide any evidence for the instrumental use of the evaluation findings. Yet our findings do not rule out the possibility that other forms of use have arisen. In general, three different forms of evaluation use are plausible. First, *process use* highlights all forms of utilisation that happened due to the process of an evaluation (Patton, 1997. However, this must not necessarily be on the basis of the recommendations, but also through the experiences they had during the evaluation. This would also explain that those school representatives who were more involved in the evaluation process also perceived a higher evaluation impact. On the other hand, the design of the IFES evaluations entails strong participation of the school representatives. They are the main data source, since they have to answer several interviews and they have to participate in focus groups. However, process use can also be instrumental, as these two types of utilisation are not mutually exclusive. Yet the source of this instrumental use is definitely not the measures that are stipulated between the schools and the Department of Education, but another form of information that has been generated during the evaluation process. Second, *conceptional use* occurs when systematic evidence challenges the policymakers' underlying assumptions and analytical concepts that determine policy choices (Weiss, 1977). The perceived evaluation quality, as well as the perceived evaluation acceptance, correlates significantly with the impact of the evaluation. It is very likely that school representatives who have a positive attitude towards the evaluation, or are at least convinced by their quality, are more likely to observe an impact, since they were informed by the evaluation. However, this would require the measurement of this specific form of use, which has not been done for this study. Third, *evaluation influence* is intended to cover the various effects of evaluations as a whole (Kirkhart, 2004; Mark and Henry, 2004). The idea of replacing evaluation use with the concept of influence is based on the critique of the historically evolved concept of use. Accordingly, the use is unilaterally focused on the results, the instrumental function and short-term effects (Kirkhart, 2004). In addition, the forms of use – such as instrumental or process-based conceptual use – can overlap (Alkin and Taut, 2003; Mark and Henry, 2004). Hence, the IFES evaluation could influence the development of the school quality management. However, this form of use requires causal inference and has hardly ever been empirically applied. Future research should focus on these forms of use in order to better understand the relationship between evaluation quality and evaluation use.

Conclusion

This article has analysed the relationship between evaluation quality and perceived evaluation use. In contrast to previous studies, we measured the quality of an evaluation externally with a systematic meta-evaluation. Our results suggest that well-defined measures stipulated by public authorities do not appear to be the core pathway of evaluation influence in the eyes of the representatives of the evaluated schools. This result is in line with a large body of literature that highlights more indirect, cognitive, and complex pathways of influence, as it suggests that there is no direct link between the implementation of evaluation findings and the impact of an evaluation (Mark and Henry, 2004; Weiss et al, 2008). In contrast, our findings suggest that analysed evaluations are not used as instrumental, but rather within the process of the evaluation, even though our empirical analysis is not able to illustrate that directly. Moreover, recent studies on external school evaluations (school inspections) show that 'accepting feedback' – schools receive, interpret, and use the evaluative information as feedback to devise and implement actions designed to improve the school – is not or is even negatively associated with reported school improvement efforts for self-evaluation, capacity building, or school effectiveness (Ehren et al, 2015; Gustaffson et al, 2015). These studies instead suggest that external school evaluation triggers school improvements by setting expectations and generating stakeholder pressure. The present analysis is based on the perceptions of stakeholders, and considers neither direct measurement of implemented measures nor school performance improvements in key areas.

These results are important, since they underline the importance of distinguishing between the various dimensions of evaluation quality. However, the findings should be treated with caution, as the study only analysed the perceived use of evaluation, which might be biased due to misreporting (Bundi et al, 2018). Moreover, the study also shows that the perceived evaluation quality might have a non-measurable effect. This could mean that the perceived impact of an evaluation depends less on the implemented measures and more on how the quality of the evaluation is perceived by the involved stakeholders. Finally, the quality assessed by the meta-evaluation in this study is limited to the subset of evaluation standards that were selected. This is not only a result of the high standardisation of the evaluation studies, but also because we did not have enough information in order to assess the consideration of each single evaluation standard. The limited variation and scope of our analysis restricts the generalisability of our findings and potentially influences our assessment; it is difficult to say in which direction the results are influenced.

What are the implications of these findings beyond the case at hand? First, our findings emphasise distinguishing between the implementation of evaluation findings and the evaluation impact. In other words, evaluators often focus on the instrumental form of use, but they should not ignore other forms of evaluation use and maybe try to maximise these utilisation forms in the design of their evaluation (for example, through stakeholder participation). Second, the findings suggest that evaluators should be more active in advising stakeholders when it comes to evaluation use, for example, through policy narratives (Rickinson et al, 2019). We can deduct from the findings that it is not only important to carry out a high-quality evaluation, but that it is also very important to adequately prepare school actors for evaluation and evidence use. Even though an evaluation can be perceived objectively to be of a high quality, this

does not mean that this evaluation is used. Third, the findings suggest careful thought is needed about the measurement of evaluation quality and evaluation consequences in research on evaluation. As a consequence, these results have several implications for evaluation practice and practitioners. Perceived evaluation quality is important for the perceived impact of an evaluation, so evaluations should focus on specific quality standards that enhance evaluation use. Hence, evaluators have to be aware that a *systematically assessed* quality of an evaluation does not go hand in hand with the *perceived* quality of that evaluation.

Research ethics statement

The present study was realised within the framework of the evaluation of IFES on behalf of the Education Directorate of the Canton of Zurich and supported by an accompanying group. The accompanying group consisted of persons from the upper secondary schools, the education directorate (secondary school and vocational training office and education planning) as well as academic educational research, who monitored the processes within the project. Participants (such as respondents and interviewees) were informed about the project and the data analysis procedure before data collection. All information was treated confidentially and the evaluation of the data does not allow any conclusions to be drawn about individual persons. Following the survey, we prepared a report of the results, which we made available to the respondents (Widmer et al, 2015).

Notes

[1] In this article, we define an evaluation as the use of scientific procedures to systematically investigate the effectiveness or efficiency of public interventions that is adapted to their political and organisational environments and designed to inform actors in ways that improve social conditions (Rossi et al, 2018: 2).

[2] An evaluation can contribute to the success of the curriculum at three levels: First, student achievements, second, evaluation of educational programmes and materials and third, evaluation of schools. We will focus on the latter in this article.

[3] The impact of the evaluation is defined as the changes caused by the evaluation occurring among the direct and beyond the direct addressees. A list of items regarding the evaluation impact can be found in Table 4 of the Appendix.

[4] IFES is an institute of the Swiss Conference of Cantonal Ministers of Education and an agency associated with the University of Zurich.

[5] We have decided to use the SEVAL Standards in our meta-evaluation, because they allow a comprehensive and balanced assessment in addition to their methodologically open approach, because of their broad definition of quality, and because they have a high acceptance and credibility in the field of evaluation.

[6] Table A1 in the Appendix presents an overview of the meta-evaluation criteria.

[7] The criteria are defined as maximal demands and not as minimal requirements. In practice, it is difficult (sometimes even impossible) to fully meet all the criteria. However, it can be expected that the evaluation will offer explanations if a criterion is disregarded. If a criterion was observed or explained why this was not possible, the criterion was classified as considered.

[8] Krippendorf's $\alpha > 90\%$.

[9] See Table A2 in the Appendix for the index configuration.

[10] Krippendorf's α >87% and significant on the 99% level.

[11] The primary and secondary levels and the evaluation mechanisms used at both levels have converged in recent decades and have several parallels. Since there are no comparable survey instruments for the secondary level, the application is reasonable in the light of this convergence. Nevertheless, the existing differences must be taken into account in the interpretation.

[12] All indices are all averaged and equally weighted. Table 4 in the Appendix shows an overview of the operationalisation.

[13] We also checked for a time trend in assessed quality in the meta-evaluation and can exclude a cohort effect.

[14] The dependent variables were the following variables: perceived quality (Model 1); implementation (Model 2); and impact of the evaluation (Model 3).

Funding

This study did not receive any additional funding.

Contributor statement

PB conducted the data analysis, and drafted both the discussion and findings; KF wrote the introduction and the theoretical chapter; TW authored the literature review and contributed to the theoretical and analytical parts. All authors commented and contributed to all chapters.

Acknowledgements

Previous versions of this paper were presented at the AEA Evaluation conference 2016 in Atlanta, GA, US, at the EES conference 2016 in Maastricht, The Netherlands, and at the DeGEval conference 2016 in Salzburg, Austria. The authors thank all the participants for their feedback. In addition, the authors thank Nadja Rüegg, Cornelia Stadter and Jeffrey Stein for their outstanding research assistance and John Bendix for his valuable comments. We are grateful to Zachary Neal as well as the three anonymous reviewers for their valuable remarks.

Conflict of interest statement

The authors declare that there is no conflict of interest.

References

Alkin, M.C. and King, J.A. (2016) The historical development of evaluation use, *American Journal of Evaluation*, 37(4): 568–79. doi: 10.1177/1098214016665164

Alkin, M.C. and Taut, S.M. (2003) Unbundling evaluation use, *Studies in Educational Evaluation*, 29(1): 1–12. doi: 10.1016/S0191-491X(03)90001-0

Ayers, T.D. (1987) Stakeholders as partners in evaluation: a stakeholder-collaborative approach, *Evaluation and Program Planning*, 10(3): 263–71. doi: 10.1016/0149-7189(87)90038-3

Böhm-Kasper, O., Selders, M. and Lambrecht, M. (2016) Schulinspektion und Schulentwicklung: Ergebnisse der quantitativen Schulleitungsbefragung, in Arbeitsgruppe Schulinspektion (eds) *Schulinspektion als Steuerungsimpuls?*, Wiesbaden: Springer, pp 1–50, doi: https://doi.org/10.1007/978-3-658-10872-4_1.

Böttcher, W. and Keune, M. (2010) Funktionen und Effekte der Schulinspektion, in W. Böttcher, J.N. Dicke and N. Hogrebe (eds) *Evaluation, Bildung und Gesellschaft*, Münster: Waxmann, pp 151–64.

Bundi, P. (2016) What do we know about the demand for evaluation? Insights from the parliamentary arena, *American Journal of Evaluation*, 37(4): 522–41. doi: 10.1177/1098214015621788

Bundi, P., Varone, F., Gava, R. and Widmer, T. (2018) Self-selection and misreporting in legislative surveys, *Political Science Research and Methods*, 6(4): 771–89. doi: 10.1017/psrm.2016.35

Bustelo, M. (2002) *Metaevaluation as a Tool for the Improvement and Development of the Evaluation Function in Public Administrations*, Paper presented at the 5th European Evaluation Society Biennial Conference, Seville, Spain.

Cook, T.D. and Gruder, C.L. (1978) Metaevaluation research, *Evaluation Quarterly*, 2(1): 5–51. doi: 10.1177/0193841X7800200101

Cook, T.D. and Shadish, W.R. (1982) Metaevaluation: an assessment of the congressionally mandated evaluation system for community mental health centers, in G.J. Stahler and W.R. Tash (eds) *Innovative Approaches to Mental Health Evaluation*, New York: Academic Press, pp 221–53.

Cooksy, L.J. and Caracelli, V.J. (2009) Metaevaluation in practice: selection and application of criteria, *Journal of MultiDisciplinary Evaluation*, 6(11): 1–15.

Cousins, J.B. (1995) Assessing program needs using participatory evaluation: a comparison of high and marginal success cases, in J.B. Cousins and L.M. Earl (eds) *Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning*, London: Routledge, pp 55–71.

Cousins, J.B. and Leithwood, K.A. (1986) Current empirical research on evaluation utilization, *Review of Educational Research*, 56: 331–65. doi: 10.3102/00346543056003331

Cousins, J.B. and Leithwood, K.A. (1992) Enhancing knowledge utilization as a strategy for school improvement, *Knowledge: Creation, Diffusion, Utilization*, 14(3): 305–33.

Cronbach, L.J. (2000) Course improvement through evaluation, in D.L. Stufflebeam, G.F. Madaus and T. Kellaghan (eds) *Evaluation Models: Evaluation in Education and Human Services*, Dordrecht: Springer.

Daigneault, P.M. (2014) Taking stock of four decades of quantitative research on stakeholder participation and evaluation use: a systematic map, *Evaluation and Program Planning*, 45(2): 171–81. doi: 10.1016/j.evalprogplan.2014.04.003

Davies, H.T.O., Nutley, S.M. and Smith, P.C. (2000) *What Works? Evidence-based Policy and Practice in Public Services*, Bristol: Policy Press.

Dedering, K. and Mueller, S. (2011) School improvement through inspections? First empirical insights from Germany, *Journal of Educational Change*, 12(3): 301–22. doi: 10.1007/s10833-010-9151-9

Donaldson, M.L., Woulfin, S., LeChasseur, K. and Cobb, C.D. (2016) The structure and substance of teachers' opportunities to learn about teacher evaluation reform: promise or pitfall for equity?, *Equity and Excellence in Education*, 49(2): 183–201. doi: 10.1080/10665684.2016.1144831

Earl, L.M. (1995) District-wide evaluation of school improvement: a system partners approach, in J.B. Cousins and L.M. Earl (eds) *Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning*, London: Routledge, pp 21–32.

Ehren, M.C.M. and Visscher, A.J. (2006) Towards a theory on the impact of school inspections, *British Journal of Educational Studies*, 54(1): 51–72. doi: 10.1111/j.1467-8527.2006.00333.x

Ehren, M.C.M., Gustafsson, J.E., Altrichter, H., Skedsmo, G., Kemethofer, D. and Huber, S.G. (2015) Comparing effects and side effects of different school inspection systems across Europe, *Comparative Education*, 51(3): 375–400. doi: 10.1080/03050068.2015.1045769

Fleischer, D.N. and Christie, C.A. (2009) Evaluation use: results from a survey of U.S. American evaluation association members, *American Journal of Evaluation*, 30(2): 158–75. doi: 10.1177/1098214008331009

Frey, K. (2010) Revising road safety policy: the role of systematic evidence in Switzerland, *Governance*, 23(4): 667–90. doi: 10.1111/j.1468-0491.2010.01503.x

Gaertner, H., Wurster, S. and Pant, H.A. (2014) The effect of school inspections on school improvement, *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, 25(4): 489–508. doi: 10.1080/09243453.2013.811089

Gustafsson, J.E., Ehren, M.C.M., Conyngham, G., McNamara, G., Altrichter, H. and O'Hara, J. (2015) From inspection to quality: ways in which school inspection influences change in schools, *Studies in Educational Evaluation*, 47(1): 47–57. doi: 10.1016/j.stueduc.2015.07.002

Hatry, H.P. (1980) Pitfalls of evaluation, in G. Majone and E.S. Quade (eds) *Pitfalls of Analysis*, Chichester: John Wiley, pp 159–78.

Head, B.W. (2010) Reconsidering evidence-based policy: key issues and challenges, *Policy and Society*, 29(2): 77–94. doi: 10.1016/j.polsoc.2010.03.001

Henry, G.T. and Mark, M.M. (2003) Beyond use: understanding evaluation's influence on attitudes and actions, *American Journal of Evaluation*, 24(3): 293–314.

House, E.R. (1980) *Evaluating with Validity*, Beverly Hills, CA: Sage.

House, E.R. (1993) *Professional Evaluation: Social Impact and Political Consequences*, Beverly Hills, CA: Sage.

Husfeldt, V. (2011) Wirkungen und Wirksamkeit der externen Schulevaluation. Überblick zum Stand der Forschung, *Zeitschrift für Erziehungswissenschaft*, 14(2): 259–82.

Jacobson, M.R. and Azzam, T. (2016) Methodological credibility: an empirical investigation of the public's perception of evaluation findings and methods, *Evaluation Review*, 40(1): 29–60. doi: 10.1177/0193841X16657728

Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F. and Volkov, B. (2009) Research on evaluation use: a review of the empirical literature from 1986 to 2005, *American Journal of Evaluation*, 30(3): 377–410. doi: 10.1177/1098214009341660

Joint Committee on Standards for Educational Evaluation (1981) *Standards for Evaluations of Educational Programs, Projects, and Materials*, New York: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation (1994) *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs*, 2nd edn, Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation (2011) *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*, 3rd edn, Los Angeles, CA: Sage.

King, J. A., and Alkin, M. C. (2019) The centrality of use: Theories of evaluation use and influence and thoughts on the first 50 years of use research, *American Journal of Evaluation*, 40(3): 431–58.

Kirkhart, K.E. (2004) Reconceptualizing evaluation use: an integrated theory of influence, *New Directions for Evaluation*, 2000(88): 5–23, doi: 10.1002/ev.1188.

Lafleur, C. (1995) A participatory approach to district-level program evaluation: the dynamics of internal evaluation, in J.B. Cousins and L.M. Earl (eds) *Participatory Evaluation in Education: Studies in Evaluation Use and Organizational Learning*, London: Routledge, pp 33–54.

Ledermann, S. (2012) Exploring the necessary conditions for evaluation use in program change, *American Journal of Evaluation*, 33(2): 159–78. doi: 10.1177/1098214011411573

Lee, L.E. and Cousins, J.B. (1995) Participation in evaluation of funded school improvement: effects and supporting conditions, in J.B. Cousins and L.M. Earl (eds) *Participatory Evaluation in Education: Studies in Evaluation use and Organizational Learning*, London: Routledge, pp 72–85.

Leithwood, K.A. (1986) *The Role of the Secondary School Principal in Policy Implementation and School Improvement. Publication Sales,* The Ontario Institute for Studies in Education, 252 Bloor Street West, Toronto, Ontario M5S 1V6 Canada.

Mark, M.M. and Henry, G.T. (2004) The mechanisms and outcomes of evaluation influence, *Evaluation*, 10(1): 35–57. doi: 10.1177/1356389004042326

Marsh, D.D. and Glassick, J.M. (1988) Knowledge utilization in evaluation efforts: the role of recommendations, *Knowledge*, 9(3): 323–41. doi: 10.1177/107554708800900301

Mayring, P. (2015) *Qualitative Inhaltsanalyse: Grundlagen und Techniken*, 12. Auflage, Weinheim: Beltz.

Miller, R.L. (2015) How people judge the credibility of information: lessons for cognitive and information sciences, in S.I. Donaldson, C.A. Christie and M.M. Mark (eds) *Credible and Actionable Evidence: The Foundation for Rigorous and Influential Evaluations*, Thousand Oaks, CA: Sage, pp 39–62.

Newman, D.L., Brown, R.D. and Rivers, L. (1987) Factors influencing the decision-making process: an examination of the effect of contextual variables, *Studies in Educational Evaluation*, 13(2): 199–209. doi: 10.1016/S0191-491X(87)80034-2

Nutley, S., Walter, I. and Davies, H.T.O. (2003) From knowing to doing: a framework for understanding the evidence-into-practice agenda, *Evaluation*, 9(2): 125–48. doi: 10.1177/1356389003009002002

Odom, S.L., Brantlinger, E., Gersten, R., Horner, R.H., Thompson, B. and Harris, K.R. (2005) Research in special education: scientific methods and evidence-based practices, *Exceptional Children*, 71(2): 137–48. doi: 10.1177/001440290507100201

Oliver, K., Lorenc, T. and Tinkler, J. (2020) Evaluating unintended consequences: New insights into solving practical, ethical and political challenges of evaluation, *Evaluation*, 26(1): 61–75.

Olsen, R.B., Orr, L.L., Bell, S.H. and Stuart, E.A. (2013) External validity in policy evaluations that choose sites purposively, *Journal of Policy Analysis and Management*, 32(1): 107–21. doi: 10.1002/pam.21660

Ouimet, M., Landry, R., Ziam, S. and Bédard, P. O. (2009) The absorption of research knowledge by public civil servants, *Evidence & Policy: A Journal of Research, Debate and Practice*, 5(4): 331–50.

Patton, M.Q. (1997) *Utilization-focused Evaluation: The New Century Text*, 3rd edn, Thousand Oaks, CA: Sage.

Pawson, R. (2006) *Evidence-based Policy: A Realist Perspective*, London: Sage.

Penninckx, M., Vanhoof, J., De Maeyer, S. and Van Petegem, P. (2016) Explaining effects and side effects of school inspections: a path analysis, *School Effectiveness and School Improvement*, 27(3): 333–47. doi: 10.1080/09243453.2015.1085421

Petrosino, A., Petrosino, C.T. and Buehler, J. (2005) 'Scared Straight' and other juvenile awareness programs for preventing juvenile delinquency, *Campbell Systematic Reviews*, 1(1): 1–62.

Potts, S.A.K. (1998) *Impact of Mixed Method Designs on Knowledge Gain, Credibility, and Utility of Program Evaluation Findings*, Phoenix: Arizona State University.

Quesel, C., Husfeldt, V. and Bauer, F.D. (2011a) *Externe Schulevaluation aus der Sicht von Lehrpersonen, Schulleitungen und lokalen Schulbehörde: Eine explorative Untersuchung zur Education Governance im Kanton Aargau*, Aarau: Fachhochschule Nordwest Schweiz.

Quesel, C., Husfeldt, V., Landwehr, N. and Steiner, P. (2011b) *Wirkungen und Wirksamkeit der externen Schulevaluation*, Bern: h.e.p.-Verlag.

Rickinson, M., Sharples, J. and Lovell, O. (2020) Towards a better understanding of quality of evidence use, in S. Gorard (ed) *Getting Evidence into Education*, Abingdon: Routledge, pp 219–33.

Rickinson, M., Walsh, L., De Bruin, K. and Hall, M. (2019) Understanding evidence use within education policy: a policy narrative perspective, *Evidence & Policy*, 15(2): 235–52.

Rogers, W. (1993) Quantile regression standard errors, *Stata Technical Bulletin*, 2(9): 1–28.

Rossi, P.H., Lipsey, M.W. and Henry, G.T. (2018) *Evaluation: A Systematic Approach*, Thousand Oaks, CA: Sage.

Russon, C. and Russon, G. (eds) (2004) International perspectives on evaluation standards, *New Directions for Evaluation*, San Francisco: Jossey-Bass, 104.

Sanders, W.L. and Horn, S.P. (1998) Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: implications for educational evaluation and research, *Journal of Personnel Evaluation in Education*, 12(3): 247–56. doi: 10.1023/A:1008067210518

Schwartz, R. and Mayne, J. (2005) Assuring the quality of evaluative information: theory and practice, *Evaluation and Program Planning*, 28(1): 1–14. doi: 10.1016/j.evalprogplan.2004.10.001

Scriven, M. (1969) An introduction to meta-evaluation, *Educational Product Report*, 2(5): 36–8.

Seidel, T., Mok, S.Y., Hetmanek, A. and Knogler, M. (2017) Meta-Analysen zur Unterrichtsforschung und ihr Beitrag für die Realisierung eines Clearing House Unterricht für die Lehrerbildung, *Zeitschrift für Bildungsforschung*, 7(3): 311–25. doi: 10.1007/s35834-017-0191-6

Slavin, R.E. (2008) Perspectives on evidence-based research in education: what works? Issues in synthesizing educational program evaluations, *Educational Researcher*, 37(1): 5–14. doi: 10.3102/0013189X08314117

Steenbergen, M.R. and Jones, B.S. (2002) Modeling multilevel data structures, *American Journal of Political Science*, 46(1):218–37. doi: 10.2307/3088424

Stufflebeam, D.L. (1974) Meta-evaluation, *Evaluation Center Occasional Paper Series #3*, Kalamazoo, MI: Western Michigan University.

Stufflebeam, D.L. (2001) The metaevaluation imperative, *American Journal of Evaluation*, 22(2): 183–209. doi: 10.1177/109821400102200204

Turnbull, B. (1999) The mediating effect of participation efficacy on evaluation use, *Evaluation and program Planning*, 22(2): 131–40.

Verhoest, K., Verschuere, B. and Bouckaert, G. (2007) Pressure, legitimacy, and innovative behavior by public organizations, *Governance*, 20(3): 469–97. doi: 10.1111/j.1468-0491.2007.00367.x

Weiss, C.H. (1977) Research for policy's sake: the enlightenment function of social research, *Policy Analysis*, 3(4): 531–45.

Weiss, C.H. (1998) Have we learned anything new about the use of evaluation?, *American Journal of Evaluation*, 19(1): 21–33. doi: 10.1177/109821409801900103

Weiss, C.H. and Bucuvalas, M.J. (1980) Truth tests and utility tests: decision-makers' frames of reference for social science research, *American Sociological Review*, 45(2): 302–13. doi: 10.2307/2095127

Weiss, C.H., Murphy-Graham, E. and Birkeland, S. (2005) An alternate route to policy influence: how evaluations affect DARE, *American Journal of Evaluation*, 26(1): 12–30. doi: 10.1177/1098214004273337

Weiss, C.H., Murphy-Graham, E., Petrosino, A. and Gandhi, A.G. (2008) The fairy godmother – and her warts: making the dream of evidence-based policy come true, *American Journal of Evaluation*, 29(1): 29–47. doi: 10.1177/1098214007313742

Widmer, T. (1996) *Meta-Evaluation: Kriterien zur Bewertung von Evaluationen*, Bern: Haupt.

Widmer, T. (2004) The development and status of evaluation standards in Western Europe, *New Directions for Evaluation*, 2004(104): 31–42. doi: 10.1002/ev.134

Widmer, T. (2005) Instruments and procedures for assuring evaluation quality: a Swiss perspective, in R. Schwartz and J. Mayne (eds) *Quality Matters. Seeking Confidence in Evaluating, Auditing and Performance Reporting*, New Brunswick: Transaction Publishers, pp 41–68.

Widmer, T., Landert, C. and Bachmann, N. (2000) *Evaluations-Standards der Schweizerischen Evaluationsgesellschaft (SEVAL-Standards)*, Bern: SEVAL.

Widmer, T., Frey, K., Rüegg, N., Stadter, C., Bundi, P. and Stein, J. (2015) Qualität der IFES-Schulevaluationen und deren Nutzung im Kanton Zürich, *Zürcher Politik- & Evaluationsstudien Nr. 13*, Zürich: Institut für Politikwissenschaft.

Appendix

Table A1: Overview of external evaluation quality assessment criteria

Standard	Description	Included
Utility The utility standards guarantee that an evaluation is oriented to the information needs of the intended users of the evaluation.		
Identifying stakeholders	Those persons participating in, and affected by, an evaluation are identified in order that their interests and needs can be taken into account.	Yes
Clarifying the objectives of the evaluation	All who are involved in an evaluation will ensure that the objectives of the evaluation are clear to all stakeholders.	No
Evaluator credibility	Those who conduct evaluations are both competent and trustworthy; this will help ensure the results an evaluation reaches are accorded the highest degree of acceptance and credibility possible.	No
Scope and selection of information	The scope and selection of the information that has been collected makes it possible to ask pertinent questions about the object of the evaluation. Such scope and selection also takes into account the interests and needs of the parties commissioning the evaluation, as well as other stakeholders.	No*
Transparency of value judgements	The underlying reasoning and points of view upon which an interpretation of evaluation results rests are described in such a manner that the bases for the value judgments are clear.	Yes
Clarity in reporting	Evaluation reports describe the object of evaluation – including its context, goals, questions posed, and procedures used, as well as the findings reached in the evaluation – in such a manner that the most pertinent information is available and readily comprehensible.	Yes
Timely reporting	Significant interim results, as well as final reports, are made available to the intended users such that they can be utilised in a timely manner.	No*
Evaluation impact	The planning, execution, and presentation of an evaluation encourage stakeholders both to follow the evaluation process and to use the evaluation.	No
Feasibility The feasibility standards ensure that an evaluation is conducted in a realistic, well-considered, diplomatic and cost-conscious manner.		
Practical procedures	Evaluation procedures are designed such that the information needed is collected without unduly disrupting the object of the evaluation or the evaluation itself.	No*
Anticipating political viability	The various positions of the different interests involved are taken into account in planning and carrying out an evaluation in order to win their cooperation and discourage possible efforts by one or another group to limit evaluation activities or distort or misuse the results.	No
Cost effectiveness	Evaluations produce information of a value that justifies the cost of producing them.	No

(*Continued*)

Table A1: (*Continued*)

Standard	Description	Included
Propriety The propriety standards ensure that an evaluation is carried out in a legal and ethical manner and that the welfare of the stakeholders is given due attention.		
Formal written agreement	The duties of the parties who agree to conduct an evaluation (specifying what, how, by whom, and when what is to be done) are set forth in a written agreement in order to obligate the contracting parties to fulfill all the agreed upon conditions, or if not, to renegotiate the agreement.	No
Ensuring individual rights and wellbeing	Evaluations are planned and executed in such a manner as to protect and respect the rights and wellbeing of individuals.	No
Respecting human dignity	Evaluations are structured in such a manner that the contacts between participants are marked by mutual respect.	No
Complete and balanced assessment	Evaluations are complete and balanced when they assess and present the strengths and weaknesses that exist in the object being evaluated, in a manner that strengths can be built upon and problem areas addressed.	Yes
Making findings available	The parties who contract to an evaluation ensure that its results are made available to all potentially affected persons, as well as to all others who have a legitimate claim to receive them.	No*
Declaring conflicts of interest	Conflicts of interest are addressed openly and honestly so that they compromise the evaluation process and conclusions as little as possible.	No
Accuracy The accuracy standards ensure that an evaluation produces and disseminates valid and usable information.		
Precise description of the object of evaluation	The object of an evaluation is to be clearly and precisely described, documented, and unambiguously identified.	Yes
Analysing the context	The influences of the context on the object of evaluation are identified.	Yes
Precise description of goals, questions, and procedures	The goals pursued, questions asked, and procedures used in the evaluation are sufficiently precisely described and documented that they can be identified as well as assessed.	Yes
Trustworthy sources of information	The sources of information used in an evaluation are sufficiently precisely described that their adequacy can be assessed.	No
Valid and reliable information	To ensure the validity and reliability of the interpretation, it is necessary to select, develop, and employ procedures for that given purpose.	No
Systematic checking for errors	The information collected, analysed, and presented in an evaluation is systematically checked for errors.	No
Qualitative and quantitative analysis	Qualitative and quantitative information are systematically and appropriately analysed in an evaluation, in a manner that the questions posed by the evaluation can actually be answered.	Yes

Table A1: (*Continued*)

Standard	Description	Included
Substantiated conclusions	The conclusions reached in an evaluation are explicitly substantiated in such a manner that stakeholders can comprehend and judge them.	Yes
Neutral reporting	Reporting is free from distortion through personal feelings or preferences on the part of any party to the evaluation; evaluation reports present conclusions in a neutral manner.	Yes
Meta-evaluation	The evaluation itself is evaluated on the basis of existing (or other relevant) standards such that the evaluation is appropriately executed, and so that stakeholders can, in the end, assess the evaluation's strengths and weaknesses.	No
Zurich		
Thematic coverage	The IFES SE includes a meta-evaluation of the quality management of school quality across all areas and may additionally contain an evaluation of a focus topic chosen jointly by the school and the public authorities.	No*
Coverage of central questions	The IFES SE answers specified central questions for the external evaluation sec II, which are defined by the canton of Zurich.	Yes
IFES		
Departure from SEVAL standards	The IFES SE contains an explicit presentation and a convincing justification when there is a deviation from the SEVAL Standards.	No*
Exogenous	The IFES SE is exogenous.	No*
Basic design	The basic design is considered adequately in the IFES SE.	No*
Need orientation	The IFES SE considers the needs of schools, cantons and the federal government.	Yes
Fit	The IFES SE is adapted to the individual school and its development.	No*
Triangulation	Various stakeholders are interviewed during the IFES SE and each topic is studied with different methods.	No
Socially competent leadership	Concerns are collected during IFES SE and discussed with the evaluation process.	No

* Excluded due to non-variance.
Source: Widmer et al, 2000

Table A2: Overview of operationalisation

Variable	Operationalization	Source: Widmer et al. 2015
Perceived evaluation quality		
Evaluator		
competence	How do you assess the competence of the IFES evaluation team? Very poor / somewhat poor / fairly good / very good	
Structure		
organisation	How do you assess the organisation of the IFES evaluation? Very poor / somewhat poor / fairly good / very good	
Process		
communication	How do you assess the communication of the IFES evaluation team? Very poor / somewhat poor / fairly good / very good	
School orientation	The IFES evaluations cater to the needs of schools Strongly disagree / partly disagree / partly agree / strongly agree	
Relevant information	The IFES evaluations collect the relevant information for the assessment of the schools. Strongly disagree / partly disagree / partly agree / strongly agree	
Opinions	The IFES evaluations take different perspectives and opinions into account. Strongly disagree / partly disagree / partly agree / strongly agree	
School portfolio	The school portfolios are adequately addressed in the IFES evaluations. Strongly disagree / partly disagree / partly agree / strongly agree	
Quality of presentation	How do you assess the quality of the oral presentations and / or meetings of the evaluation results by the IFES evaluation team? Very poor / somewhat poor / fairly good / very good	
Evaluation report		
Coverage	How do you assess the coverage of the IFES evaluation report? Very poor / somewhat poor / fairly good / very good	
Language	How do you assess the language of the evaluation report? Very poor / somewhat poor / fairly good / very good	
Layout	How would you rate the design of the evaluation report (structure, layout, and so on)? Very poor / somewhat poor / fairly good / very good	
Content	How do you assess the content of the evaluation report? Very poor / somewhat poor / fairly good / very good	
Results		
Convincing	The results of the IFES evaluation of the quality management are convincing. Strongly disagree / partly disagree / partly agree / strongly agree	
Adequate	The elements of quality management are adequately covered. Strongly disagree / partly disagree / partly agree / strongly agree	
Useful	The results of the quality management were / are useful. Strongly disagree / partly disagree / partly agree / strongly agree	
Surprising	The results of the quality management were / are (partly) surprising. Strongly disagree / partly disagree / partly agree / strongly agree	
Recommendation		
Coherent	The recommendations are clear and coherent. Strongly disagree / partly disagree / partly agree / strongly agree	

(*Continued*)

Table A2: (*Continued*)

Variable	Operationalization	Source: Widmer et al. 2015
Justified	The recommendations are well justified. Strongly disagree / partly disagree / partly agree / strongly agree	
Feasible	The recommendations are feasible. Strongly disagree / partly disagree / partly agree / strongly agree	
Helpful	The recommendations are useful. Strongly disagree / partly disagree / partly agree / strongly agree	
Perceived implementation	To what extent did your school implement the measures which your school stipulated with the authorities? Not at all / partly / to a large degree / fully	
Perceived impact	Additive index of 13 factors: The IFES evaluation has an impact on the … –… students' performance –… school atmosphere –… teaching atmosphere –… teaching quality –… working atmosphere –… contacts with parents –… public presentation –… supervision of students –… school principal –… school board –… quality management –… school regulation –… legal compliance Strong positive impact / positive impact / no impact / negative impact / strong negative impact	
Individual characteristics		
Influence	To what extent could you influence the following activities as part of the IFES evaluation? –Evaluation instruments –School attendance –Formulation of recommendations Not at all / partly / fully	
Gender	Gender of interviewed person Male; female	
School characteristics		
Year of evaluation	Year of evaluation 2006/2007–2013/2014	
School type	Type of school Secondary school and vocational school	